JASA ARTICLE

High-resolution temporal weighting of interaural time differences in speech

Lucas S Baltzell[a] and Virginia Best[b]
Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215, USA

ABSTRACT:
Previous studies have shown that for high-rate click trains and low-frequency pure tones, interaural time differences (ITDs) at the onset of stimulus contribute most strongly to the overall lateralization percept (receive the largest perceptual weight). Previous studies have also shown that when these stimuli are modulated, ITDs during the rising portion of the modulation cycle receive increased perceptual weight. Baltzell, Cho, Swaminathan, and Best [(2020). J. Acoust. Soc. Am. **147**, 3883–3894] measured perceptual weights for a pair of spoken words ("two" and "eight"), and found that word-initial phonemes receive larger weight than word-final phonemes, suggesting a "word-onset dominance" for speech. Generalizability of this conclusion was limited by a coarse temporal resolution and limited stimulus set. In the present study, temporal weighting functions (TWFs) were measured for four spoken words ("two," "eight," "six," and "nine"). Stimuli were partitioned into 30-ms bins, ITDs were applied independently to each bin, and lateralization judgements were obtained. TWFs were derived using a hierarchical regression model. Results suggest that "word-initial" onset dominance does not generalize across words and that TWFs depend in part on acoustic changes throughout the stimulus. Two model-based predictions were generated to account for observed TWFs, but neither could fully account for the perceptual data. © *2021 Acoustical Society of America*.

https://doi.org/10.1121/10.0005934

(Received 2 April 2021; revised 29 July 2021; accepted 30 July 2021; published online 23 August 2021)

[Editor: G. Christopher Stecker] Pages: 1311–1320

I. INTRODUCTION

When extracting information about the location of a sound source in the world, listeners must integrate binaural cues over time and frequency. Using controlled stimuli, numerous studies have shown that this integration is based on a non-uniform weighting of these cues (in particular interaural time differences, ITDs), and a number of distinct weighting phenomena have been observed implicating both central and peripheral mechanisms (for reviews, see Clifton and Freyman, 1997; Stecker *et al.*, 2021; Best *et al.*, 2021). The extent to which these phenomena generalize to spectrotemporally complex speech signals is not clear. In a recent attempt to characterize the perceptual weighting of ITD cues in speech, Baltzell *et al.* (2020) measured spectrotemporal weights for a pair of monosyllabic spoken words ("two" and "eight"). Using a regression model to relate ITDs in the stimulus to lateralization judgements, they found that frequency bands between approximately 400 and 1000 Hz received the largest perceptual weight, consistent with previous studies identifying a "spectral dominance" region, broadly peaked between 600 and 800 Hz. They also found that perceptual weights were largest for phonemes at the beginning of the word, consistent with previous studies demonstrating the increased weighting of cues at sound onset ("onset dominance"). This result seemed to reflect the influence of a temporal weighting mechanism that favored the global onset of the speech stimulus ("word-onset dominance"), rather than local fluctuations throughout the stimulus.

Generalizability of this conclusion suffered from two limitations. First, the words "two" and "eight" each contain salient acoustic onsets at the beginning of the word (Figs. 1 and 2). For "two" there is a word-initial broadband transient from the voiceless plosive /t/, and for "eight," there is an abrupt, high-energy onset of the vowel /eɪ/. Since onset dominance for controlled stimuli is known to depend on a number of acoustic factors including onset steepness (e.g., Klein-Hennig *et al.*, 2011), we wanted to examine words with less salient word-initial onsets. In the current study, in addition to the words "two" and "eight," we obtained temporal weighting functions (TWFs) for the words "six" and "nine." For "six," the fricative /s/ does not have a steep onset or much low-frequency energy, and so may not be as effective at carrying ITD cues (and thus driving perceptual weighting). For "nine," despite a clear word-initial voicing onset, the peak in voiced energy does not occur until the word-medial vowel /ai/, and this more gradual onset of voiced energy may also be less effective. A second limitation of the previous study was that we only obtained weights for two time bins, making it difficult to observe any influence of local temporal fluctuations throughout the word. Since previous studies have demonstrated that weighting

[a] Electronic mail: lbaltzel@bu.edu, ORCID: 0000-0002-7082-9960.
[b] ORCID: 0000-0002-5535-5736.

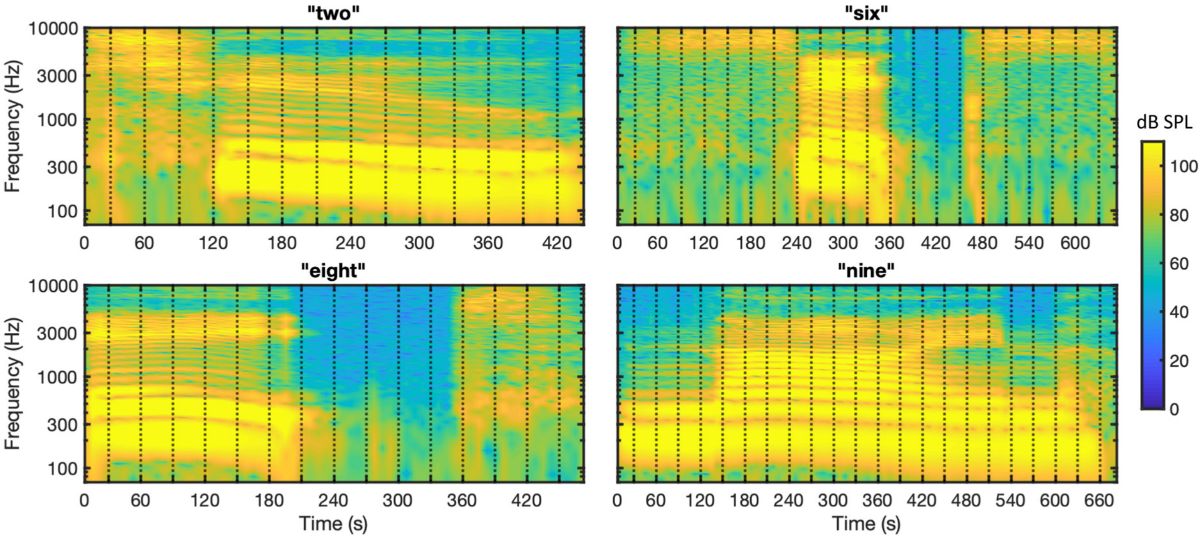

FIG. 1. (Color online) Spectrogram representation of the four speech tokens. Tokens were partitioned into 30-ms time bins, and grids overlaid on the spectrogram represent the boundaries between these bins. The overall intensity for each token was 70 dB SPL.

patterns follow envelope fluctuations for modulated sounds (e.g., Dietz *et al.*, 2013; Stecker, 2018), it is reasonable to expect that this will also be the case for speech. In the present study, we obtained TWFs with 30-ms temporal resolution. By measuring TWFs for a larger set of speech tokens and with a finer resolution, the goal of the present study was to determine (1) whether ITDs at word onset consistently receive the highest perceptual weight ("word-onset dominance"), and (2) whether perceptual weights are sensitive to envelope fluctuations throughout the word.

II. METHODS

A. Participants

Eight listeners (five female) with normal hearing (all pure-tone audiometric thresholds ≤20 dB hearing level, up to 8 kHz) between the ages of 19 to 32 (mean age = 24) participated in this study. All were native English speakers. Experiments were conducted at Boston University, and all procedures were reviewed and approved by the Institutional Review Board. All listeners provided informed consent prior to testing, and some had previous experience with psychoacoustic testing.

B. Stimuli

Four speech tokens were drawn from a corpus of monosyllabic words recorded by Sensimetrics Corporation (Malden, MA), which contained recordings from multiple male and female talkers (see Kidd *et al.*, 2008). In addition to recordings of the words "two" and "eight" (used in Baltzell *et al.*, 2020), recordings of the words "six" and

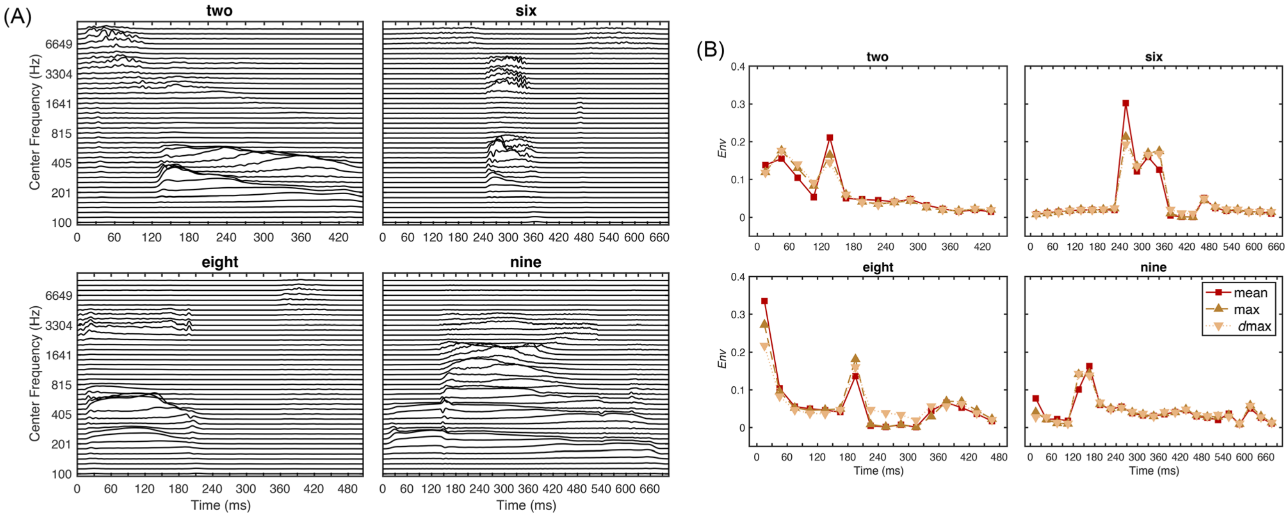

FIG. 2. (Color online) (A) Envelopes at the output of a gammatone filter bank for each word. Only odd-numbered filters (1, 3, etc.) are shown for display purposes. (B) The envelope onset metric *Env*, derived in three different ways from the envelopes in (A), are shown in arbitrary normalized units.

"nine" were also selected. The four selected words were spoken by the same female talker. Spectrograms of these word tokens are shown in Fig. 1.

Each speech stimulus was partitioned into 30-ms time bins, gated on and off with a 5-ms crossfade centered at the boundary between bins. Because the ramp and damp functions of the crossfade were mirror-image raised-cosine (1/4 cycle) windows, the sum of the windows was equal to one across all samples. Speech token durations were unequal, which resulted in 15 bins for "two," 16 bins for "eight," 22 bins for "six," and 23 bins for "nine." ITDs for each time bin were drawn independently from a ±150-μs uniform distribution (mean of 0). The same distribution was used across bins and listeners. This distribution was chosen so that the extreme ITDs would be clearly distinguishable by any listener, but not so large that variations in ITD across time bins would disrupt their binding into a single perceptual stream.[1] Within each bin, ITDs were applied in the frequency domain by taking a fast Fourier transform (FFT) of the signal in one ear (the delayed ear), shifting the phase of each frequency component corresponding to the delay specified by the ITD, and taking the IFFT of the resulting spectrum. This results in a shift of the entire waveform in the delayed ear but preserves the interaural phase relationship that would naturally occur for the desired ITD.

C. Procedure

Stimuli were presented *via* Sennheiser HD 280 headphones (Wedemark, Germany) to listeners seated in a double-walled sound-attenuating chamber (IAC Acoustics, North Aurora, IL). The digital signals were generated on a PC outside of the booth and then routed through an RME HDSP 9632 24-bit soundcard (Haimhausen, Germany). Stimulus presentation level was normalized based on headphone calibration to a flat-spectrum broadband noise, and inverse filtering based on the headphone frequency response was not applied. All stimuli were presented at 70 dB sound pressure level (SPL).

Listeners were asked to make lateralization judgments for the speech tokens with non-uniform ITDs. On each trial, two copies of the same word were presented, separated by a 500-ms inter-stimulus interval. The first copy served as a reference and was presented diotically (all time bins with zero ITD). The second copy was the target and contained a random ITD in each time bin (see Sec. II B). Listeners were instructed to indicate whether the target word was presented from the left or right relative to the reference. They were also instructed to provide an answer even if they were not sure.

Lateralization weights were obtained for each word based on a single block of trials, and the order of blocks was randomized for each listener. The number of trials in a given block was determined by the number of time bins in the word token. Specifically, 32 trials were collected for each bin, yielding 480 trials for "two," 512 trials for "eight," 704 trials for "six," and 736 trials for "nine." Total testing time was around four hours.

D. Analysis

1. Lateralization weights

Following Baltzell *et al.* (2020), lateralization weights were obtained using a Bayesian hierarchical regression model. This model allowed us to simultaneously estimate population-level weights (γ) and individual weights (β). ITDs in each time bin were related to binary lateralization judgments (0 "left" or 1 "right") using a logistic link function, so weights are in log-odds units. Our hierarchical model was implemented in R using the "brms" package (Bürkner, 2018) with default uninformative priors. Like the model described in Baltzell *et al.* (2020), this package uses a non-centered parameterization. Also, following Baltzell *et al.* (2020), individual intercepts and slopes were assumed to be independent.[2]

This model assumes that on any given trial, listeners linearly sum ITD cues in each time bin to arrive at a single laterality estimate, where the ITDs in each time bin are multiplied by the appropriate observer weight (β) prior to summation. Given the potential differences in ITD sensitivity across bins, these weights are likely influenced by sensitivity in a non-uniform fashion. Following Baltzell *et al.* (2020), population-level weights were obtained with a model that takes as input the raw data, rather than individual model weights that have been normalized. Weights should therefore be interpreted as reflecting a combination of sensitivity and effects of temporal position. Since time bins are not acoustically equivalent across words, each word was fit with a separate model.

2. Quantifying "onsets" in stimulus envelope (Env)

We hypothesized that weighting patterns may depend in part on the steepness of the rising portions of the stimulus envelope (Klein-Hennig *et al.*, 2011). To quantify this, we defined a bank of 80 gammatone filters with equal log-spaced center frequencies between 100 Hz and 10 kHz, and passed each stimulus through this filter bank (Slaney, 1993). At the output of each filter, we computed the envelope by taking the magnitude of the analytic signal and applying a low-pass filter with 128-Hz cutoff (4th order). The resulting signals are shown in Fig. 2(A).

To compute the envelope "onset" metric *Env*, we first computed a derivative for each sample of each envelope signal, resulting in a signal that describes the rate of envelope change as a function of time (positive values indicate an increasing envelope). All derivatives less than zero were set to zero (half-wave rectified), and all derivative samples corresponding to an instantaneous intensity of less than 10 dB SPL were discarded. This positive envelope derivative signal was partitioned into 30-ms time bins (corresponding to the partitioning described in Sec. II B) and three different summary statistics were derived. Assuming that lateralization judgments are based on a simple integration of onsets within each bin, we calculated the mean positive envelope derivative signal ["mean" in Fig. 2(B)]. Assuming that lateralization judgments are based on the most salient onset within each bin, we calculated the maximum positive envelope derivative signal ["max" in Fig. 2(B)]. Assuming that salience of an onset depends not only on the onset steepness but also on the "flatness" of the envelope prior to onset (Klein-Hennig *et al.*, 2011), we calculated the difference between each positive envelope derivative sample and the mean positive envelope derivative in the preceding 10 ms, and found the maximum difference in each bin ["*d*max" in Fig. 2(B)]. For each of these summary statistics, values were summed across frequency to yield a single onset metric for each time bin. These functions are normalized (divided by sum) for display. Given the high degree of similarity across these different *Env* metrics, and the conceptual simplicity of the mean, *Env* is defined as the mean positive envelope derivative function (red curve in Fig. 2) for all subsequent analyses.

The envelope onset metric *Env*, which captures changes in intensity, is partially correlated with raw intensity, a feature known to influence spectral/temporal weighting in a variety of monaural contexts (e.g., Berg, 1990; Lutfi *et al.*, 2008; Oberfeld, 2008). The available literature suggests that changes in intensity rather than intensity itself determine the salience of binaural cues (see Sec. IV A), which is why we focused primarily on *Env*. Given the numerous studies demonstrating "loudness dominance" though, we also tried to account for lateralization judgements using intensity (in dB SPL) in each bin (see supplemental Fig. 1 in the supplementary material[3]). Due in part to the limited variation in intensity over bins (compared to the variation in *Env*), lateralization judgments were better accounted for by *Env*.

It also bears mentioning that we did not apply any weights over frequency when deriving *Env*. A number of classic studies have revealed a spectral dominance region broadly peaked at 600–800 Hz (e.g., Bilsen and Raatgever, 1973), though other recent studies have reported different spectral weighting functions. Ahrens *et al.* (2020) found that ITDs at the low-frequency edge of broadband noise contributed most strongly to lateralization judgments. Using speech stimuli, Baltzell *et al.* (2020) found that in addition to a peak in the canonical dominance region, there was a second high-frequency peak, possibly driven by energy in the stimulus. Without strong *a priori* motivation for a particular set of frequency weights, we chose not to apply any.

3. Predicting weights based on the output of an auditory nerve model (AuN)

In addition to our onset metric *Env*, we derived a set of predicted weights based on the neural representation at the output of an auditory nerve (AN) model. Following Stecker (2014), we first passed our experimental stimuli through a binaural model that outputs simulated lateralization judgments. We then used a regression model to determine the extent to which ITDs in each time bin of the stimuli accounted for the simulated lateralization judgments. From this regression, we obtained simulated lateralization weights, *AuN*. Simulated lateralization judgements were obtained for 250 trials.

For each trial, an experimental stimulus was generated (Sec. II B) and passed through a phenomenological AN model (Bruce *et al.*, 2018; see also Zilany *et al.*, 2009). We used a bank of 40 simulated AN channels, with equal log-spacing between 200 Hz and 10 kHz. From the spike trains at the output of the AN model, we computed a shuffled cross-correlogram (SCC; see Louage *et al.*, 2004) using 35 fibers from each channel. The SCC can be thought of as a neural cross correlation function, and so reflects a delay-line architecture classically attributed to the medial superior olive (see McAlpine *et al.*, 2001). To obtain a lateralization judgment, SCCs were summed over channels, a centrality weight (Stern and Shear, 1996) was applied to the result, and the centroid was computed. This centroid is in continuous lateralization units.

In order to obtain *AuN*, we used a regression model to relate ITDs in the stimulus to the centroid output. Since the centroid is continuous rather than binary, a linear rather than logistic regression was used. Normalized *AuN*-weights are shown in Fig. 3. Since neural adaptation beyond the AN is not included in this metric, nor are any binaural integration windows, *AuN* reflects the prediction if TWFs follow the representation of interaural differences in AN firing patterns.

4. Classification performance

Classification performance was assessed for four weighting functions: β, γ, *Env*, and *AuN*. The goal was to determine how well these weights could account for lateralization judgments across listeners. Rather than simply comparing the shapes of different weighting functions, we compare the extent to which these weighting functions can account for the same lateralization judgments. By conditioning the comparisons on the data in this way, we remove features of these weighting functions that do not meaningfully affect their ability to account for observed lateralization judgments. For instance, when comparing classification performance for model prediction weights *Env* and *AuN* against regression weights γ, it is possible that despite visual differences, classification performance could be essentially equal. For ease of explanation, we will first illustrate the method using β-weights. For each listener i, predicted responses $\hat{y}_i$ were generated using the equation $\hat{y}_i = logit^{-1}(\beta_i X_i)$. A receiver operating characteristic (ROC) analysis was then performed comparing predicted to actual responses, and classification accuracy is reported as area under the ROC curve (AUC). AUC provides a scale-invariant measure of performance that integrates over all classification thresholds ("classification threshold" is analogous to "response criterion"), and so is a measure of sensitivity akin to d'. In this case, classification threshold refers to lateralization bias (intercept in the regression model), which means that the bias term is irrelevant to classification performance, as are any (positive) scaling factors applied to the weight functions.

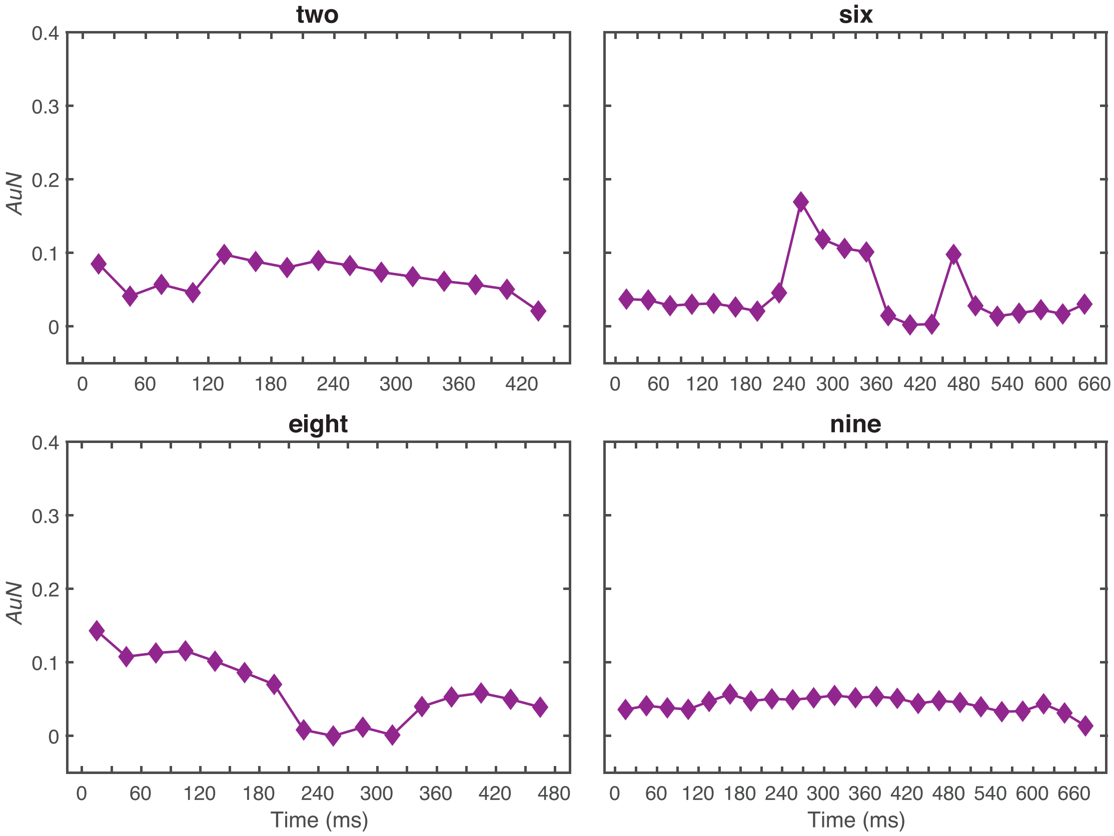

FIG. 3. (Color online) Normalized lateralization weights obtained using simulated auditory nerve responses. Error bars reflecting the standard error of the weight estimates are negligible enough as to not appear in this plot.

AUC units can be interpreted as the probability that the predicted response $\hat{y}$ for a randomly drawn right response (1) is larger than $\hat{y}$ for a randomly drawn left response (0).

The same method was applied for γ, *Env*, and *AuN*, except that the same weight values were used for all listeners since these weights are not listener dependent. The goal of this analysis was to determine (1) the degree to which population-level weights can account for individual performance, (2) the degree to which individual performance can be accounted for by onsets in the stimulus envelope (see Sec. II D 2), and (3) the degree to which individual performance can be accounted for by the AN representation of the stimulus (see Sec. II D 3). In order to determine the effect of the different "shapes" of our weighting functions, classification accuracy was also assessed for a set of uniform ("unshaped") weights. By allocating equal weight to each bin, uniform weighting reflects an un-weighted sum of the ITD values across bins, which is related to the average by an arbitrary proportionality constant. Classification performance for a set of uniform weights is therefore an appropriate baseline against which to assess classification performance for sets of non-uniform weights (i.e., β, γ, *Env*, *AuN*).

III. RESULTS

A. Lateralization weights

Lateralization weights were obtained using a Bayesian hierarchical regression model, fit separately for each speech token. For each time-frequency bin, a posterior distribution of β values was obtained for each listener, as well as a posterior distribution for the group-level mean γ. The medians of these distributions are shown in Fig. 4, along with 97.5% credible intervals. We take the median of the posterior distribution as our point estimate (Bayes estimate) of the parameter. Credible intervals indicate the range containing 97.5% of probability mass for the posterior distribution.

We see clear evidence of word-onset dominance for the words "two" and "eight." For these words, the first time bin has the largest weight, followed by a steep decline in weighting for the second time bin. For "two," there is a gradual decline following the second time bin that does not correspond to a phoneme boundary. For "eight," however, the weights remain relatively constant over the word-initial vowel, dropping off steeply at the phoneme boundary. We do not see word-onset dominance for the words "six" and "nine." Instead, we see a sharp increase in weight at the onset of the word-medial vowel. For "six," weights remain high for the duration of the vowel, dropping off during a silent period before the onset of the word-final consonant. The time bin containing the onset of this consonant is upweighted, resulting in a bimodal weight pattern. For "nine," weights do not remain high for the duration of the vowel. Overall weights are lower for "nine" than "six," perhaps due to the lack of clear onset of voicing. Overall weights are lower for "six" and "nine" compared to "two"

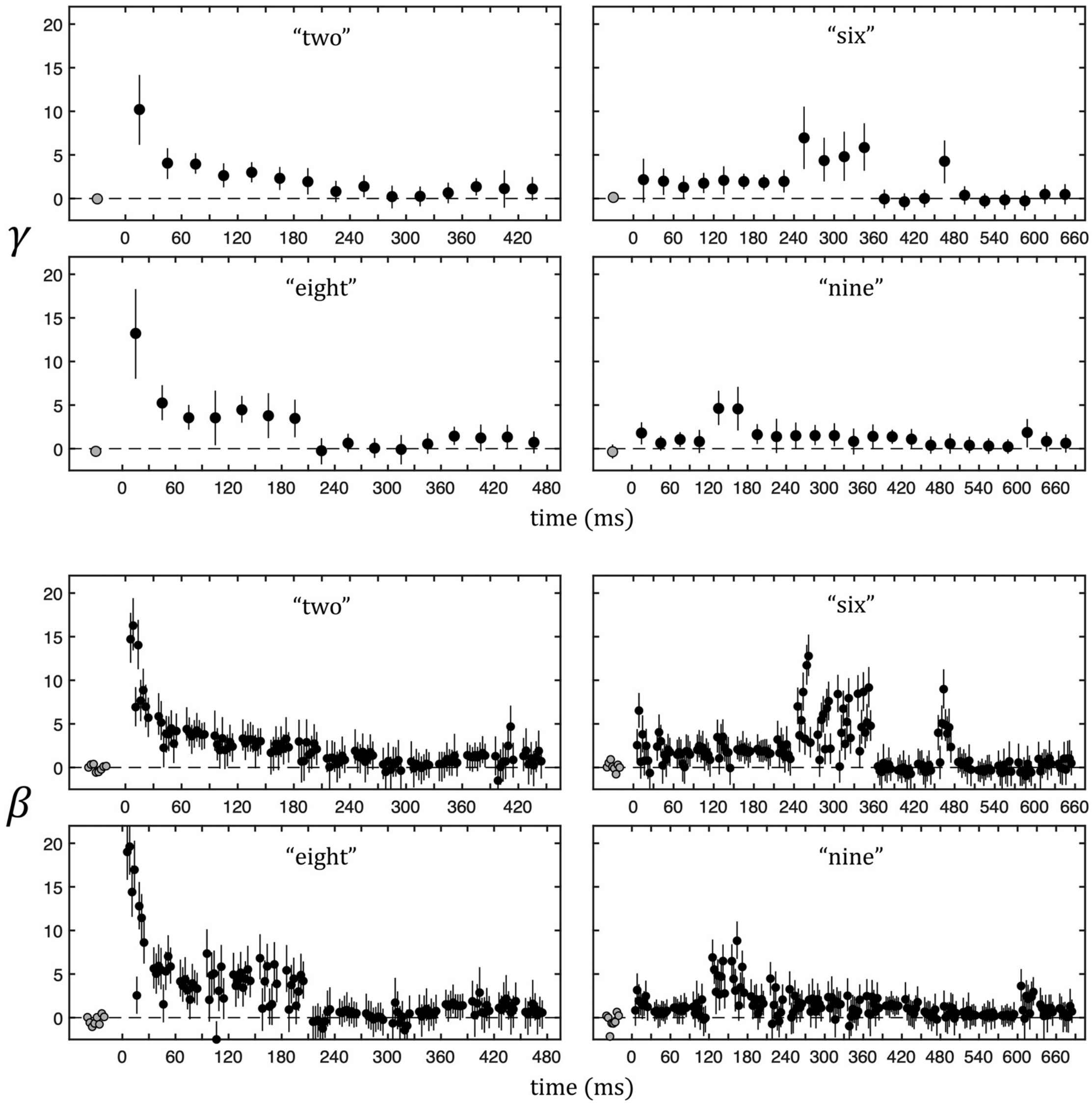

FIG. 4. Bayes estimates of population-level parameters γ and individual-level parameters β for each word token. Error bars indicate 97.5% credible intervals. The intercept terms (leftmost gray dots) reflect response bias for lateralization judgments across all time-frequency bins. The dashed line at 0 indicates the prediction of the null hypothesis that ITDs have no relationship to lateralization judgments.

and "eight," perhaps due to the lack of a strong word-onset cue. There is no evidence for an increase in weight towards the end of the words that might reflect an offset or recency effect (Stecker and Hafter, 2002).

B. Classification performance

Classification performance (AUC) was assessed for individual-level lateralization weights β, population-level lateralization weights γ, and the model predictions *Env* and *AuN*. To establish baseline performance, AUC was also assessed for a set of uniform weights. These weight functions are shown in Fig. 5(A) for each word, with each function normalized (divided by its sum) to be on the same arbitrary scale. Since AUC is immune to rescaling, this normalization has no effect on classification performance. AUC values are shown in Fig. 5(B). The uniform-weighting baseline performance (blue) reflects the extent to which a simple average across time bins can account for lateralization judgments. Classification performance for β (gray) is also shown, and since these weights reflect model parameters specific to each listener, this represents a ceiling on classification performance.

One goal of this analysis was to determine how well population-level weight estimates (γ) could account for individual lateralization judgments, and another was to determine how well these judgments can be accounted for by model predictions *Env* and *AuN*. In this case, "well" is a relative term that is bounded between ceiling performance for β and baseline performance for uniform weights. For γ, we are particularly interested in the comparison with β, as this tells us the extent to which population-level weights can account for lateralization judgments across individuals. For *Env* and *AuN*, we are particularly interested in the

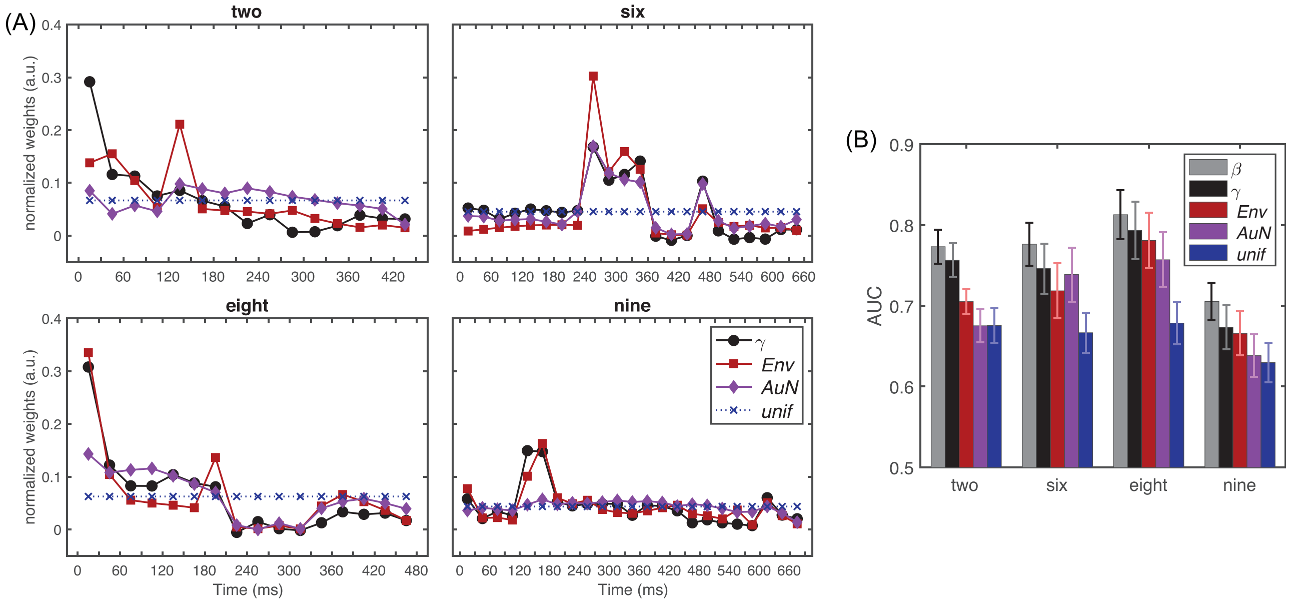

FIG. 5. (Color online) Classification performance. (A) Weights used for classification, normalized for ease of display. Also, for ease of display, individual weights β are not shown (see Fig. 3). (B) Mean classification accuracy across listeners in AUC units. Error bars indicate standard error of the mean.

comparison with γ, as this tells us the extent to which the classification performance of population-level weights can be accounted for by model predictions.

Differences in classification performance were assessed using paired *t*-tests, and significance is reported using an uncorrected two-tailed criterion of $\alpha = 0.05$. The difference in classification performance between β and γ ($\beta > \gamma$) was statistically significant for all words ("two" $p = 0.01$; "six" $p = 0.018$; "eight" $p = 0.023$; "nine" $p < 0.001$). Despite significant differences, however, classification performance tended to be similar for β and γ, consistent with the fact that temporal weighting functions were not highly variable across individuals (Fig. 4).

The difference in classification performance between γ and *Env* ($\gamma > Env$) was statistically significant for all words ("two" $p = 0.007$; "six" $p = 0.009$; "eight" $p = 0.002$; "nine" $p = 0.02$). Despite the statistically significant differences for "eight" and "nine," however, classification performance (AUC) was very similar for these words ("eight" $\gamma = 0.79$, $Env = 0.78$; "nine" $\gamma = 0.673$, $Env = 0.666$). We suggest that the population-level weights for "eight" and "nine" are largely accounted for by the envelope onset model *Env*, whereas weights for "two" and "six" are not. The difference in classification performance between *Env* and uniform was statistically significant for "six" ($p = 0.02$), "eight" ($p = 0.001$), and "nine" ($p = 0.001$), but not for "two" ($p = 0.07$).

The difference in classification performance between γ and *AuN* ($\gamma > AuN$) was statistically significant for "two" ($p = 0.01$), "eight" ($p = 0.009$) and "nine ($p = 0.003$), but not for "six" ($p = 0.09$). This suggests that population-level weights for "six" are well accounted for by the AN-based binaural model *AuN*, though this model could not account for "two," "eight," and "nine." It might be noted though that for "eight," classification performance for *AuN* is quite good (0.76), despite being significantly worse than γ (0.79). The difference in classification performance between *AuN* and uniform was statistically significant for "six" ($p = 0.002$), "eight" ($p < 0.001$), and "nine" ($p = 0.02$), but not for "two" ($p = 0.98$).

While we did not evaluate differences between *Env* and *AuN* statistically, it should be noted that in general, classification performance for *Env* was superior.

IV. DISCUSSION

The primary goal of this study was to determine whether the "word-onset" dominance observed by Baltzell *et al.* (2020) generalized to words with less salient word-initial acoustic onsets. We found that word-onset dominance does *not* generalize, and that instead, weighting patterns depend to some extent on the acoustics of each stimulus. This result calls into question the discussion of temporal effects offered by Baltzell *et al.* (2020), who suggested the influence of a central mechanism suppressing ITD cues following word onset. Instead, the present results suggest that TWFs for ITDs in speech may be largely accounted for by peripheral mechanisms that are sensitive to the details of the temporal envelope (e.g., Dietz *et al.*, 2013; Stecker, 2014). In the discussion that follows, we will consider the extent to which the TWFs we observed can be accounted for by model predictions, as well as the implications of our results for the design of binaural listening devices.

A. Do TWFs for speech follow acoustic onsets?

Using modulated high-frequency tones, Klein-Hennig *et al.* (2011) showed that ITD sensitivity strongly depended on the steepness of the rising portion and duration of the preceding silence, while the steepness of the falling portion and the duration of the peak had no effect. Similarly, Dietz *et al.* (2013) found that ITDs at the rising portion of the modulation cycle dominated the lateralization percept for modulated low-frequency stimuli. Specifically, they showed that lateralization judgments were dominated by ITDs in the low-frequency carrier (temporal fine structure, TFS) that coincided with the rising portions of amplitude-modulated binaural beats. Since the envelope was diotic (and so did not itself contain an interaural difference), their result suggested that the sampling of binaural cues from the TFS depended on the characteristics of the envelope (see also Schimmel *et al.*, 2008; Stecker and Bibee, 2014). Together, these results suggest a common influence of modulation for both high-frequency and low-frequency carriers. Rising-portion dominance for modulated sounds was also observed by Stecker (2018) using broadband carriers, and despite evidence that rising-portion dominance depends on carrier frequency (Hu *et al.*, 2017), a parsimonious account of these studies is that transient onsets in the envelope of modulated sounds trigger the sampling of binaural cues (Stecker *et al.*, 2021). This account is also consistent with the numerous studies demonstrating onset dominance for high-rate click trains (e.g., Hafter and Dye, 1983; Saberi, 1996; Stecker and Hafter, 2002), although the extent to which this "sampling" reflects active (top-down) or passive (bottom-up) processes is still unclear (see Sec. IV B).

A prediction from this literature is that TWFs will reveal large weights where there are transient onsets in temporally modulated stimuli such as speech. To quantify this prediction for our stimuli, we constructed an onset metric *Env*, based on the output of a gammatone filter bank. We found that for "eight" and "nine," this metric could effectively account for observed TWFs, suggesting a strong link between transient onsets in the envelope and the sampling of ITDs. However, we found that *Env* could not account for observed TWFs for "two" and "six," suggesting that the simple rising-slope model we constructed is insufficient to relate speech acoustics to TWFs.

B. Can TWFs for speech be predicted using the output of an auditory nerve model?

In previous studies using controlled stimuli, the interpretation of ITD TWFs was simplified by the fact that each temporal unit was acoustically identical and thus the sensitivity of listeners to the ITD in each unit (when presented in isolation) was equal. This is not necessarily true of our speech stimuli, where acoustic differences across bins may lead to variations in ITD sensitivity. In an effort to quantify peripheral representation of ITDs in each bin, we determined the extent to which ITDs in the stimulus influenced the lateral position estimate based on simulated AN responses. These simulated weights *AuN* reflect the "sensitivity" of each bin ("sensitivity" refers to the contribution of each bin to the overall inter-spike interval histogram as captured by the SCC) along with effects of neural adaptation that arise from temporal ordering.

We found that TWFs for "six" could be accounted for by *AuN*, but that TWFs for "two," "eight," and "nine" could not. As with *Env*, we found that an AN-based binaural model is insufficient to relate speech acoustics to TWFs.

C. Peripheral vs central mechanisms

While Baltzell *et al.* (2020) hypothesized a word-onset dominance for speech, the results of the present study suggest instead that weighting patterns depend in part on local onsets throughout the word. While this seems to be more consistent with a peripheral mechanism, neither *Env* nor *AuN* could fully account for observed TWFs. Both of these model predictions seem to be incomplete, and it is not obvious to us why TWFs for certain words are well accounted for by a given model, and others not. Developing a more accurate model prediction will likely require a larger dataset to prevent overfitting, and will be the focus of future research.

While our results are broadly consistent with peripheral mechanisms, there are a number of caveats worth mentioning. First, central mechanisms may be needed to explain why dynamic variations in speech TFS do not influence TWFs. Previous studies have shown that for high-rate trains of stimulus tokens, binaural sampling "resets" when a novel token is presented, or when some other irregularity/aperiodicity is introduced (e.g., Hafter and Buell, 1990; Freyman *et al.*, 1997; Brown and Stecker, 2011; Stecker, 2018). Weights are observed to be higher corresponding to these novel acoustic events, a phenomenon that has been characterized as a central breakdown of echo suppression (for a review see Clifton and Freyman, 1997). For speech, our results suggest that while transient onsets in the envelope tend to trigger the sampling of binaural cues, various dynamic changes in the TFS do not (consistent with the results of Dietz *et al.*, 2013; Fig. 1).[4] This result may implicate a central mechanism responsible for making ongoing predictions about the stimulus. To determine the extent to which top-down acoustic expectations influence temporal weighting of spectro-temporally complex sounds, it will be useful to obtain weights using unfamiliar non-speech stimuli, or to use paradigms where the acoustic stimulus is not presented repeatedly over the course of the experiment.

A second caveat concerns the time scales of binaural changes in our stimuli. Previous studies have shown that temporal weighting effects break down for unmodulated click trains when the inter-click interval (ICI) becomes too large. Stecker (2014) showed that for click trains with inter-click intervals less than 10 ms, both TWFs and peripheral model-based predictions revealed onset (of click train) dominance and recency effects. However, for click trains with 10-ms ICIs, both TWFs and peripheral model-based predictions were flat. This is consistent with the time course of neural adaptation in the AN, which is largely released after 10 ms. It is also consistent with the operating range of the precedence effect for click pairs (also ∼10 ms), which is thought to be similarly influenced by neural adaptation in the periphery (see Stecker & Hafter, 2002; Litovsky *et al.*, 1999; Brown *et al.*, 2015). Since our stimuli were partitioned into 30-ms bins, we might expect the influence of peripheral neural adaptation from one bin to another to be relatively small. The fact that *AuN* tends to follow modulation contours rather than revealing simple onset/recency effects is broadly consistent with this expectation. On the other hand, the operating range for the precedence effect is much larger for speech than for clicks, and can be observed out to around 50 ms (e.g., Lochner and Burger, 1958), perhaps suggesting an influence of more central mechanisms or sources of adaptation (e.g., Fitzpatrick *et al.*, 1999; van der Heijden *et al.*, 2019). Central mechanisms likely influence binaural integration windows, which can vary widely in size depending on both task and stimulus (Culling and Colburn, 2000; Dietz *et al.*, 2011; Hauth and Brand, 2018), and may help explain why TWFs (γ) were not well accounted for by *AuN*.

D. Practical considerations for binaural devices

The speech envelope has long been the focus of speech enhancement algorithms, designed to improve speech intelligibility in various listening environments (e.g., Lorenzi *et al.*, 1999; Apoux *et al.*, 2001; Apoux *et al.*, 2004; Koning and Wouters, 2012; Desloge *et al.*, 2017). We are also aware of at least two studies that have attempted to provide binaural benefit through manipulation of the temporal envelope. Francart *et al.* (2014) found improved ITD sensitivity for bimodal cochlear implant (CI) users when stimulation in the implanted ear was modulated such that it corresponded explicitly to F0-related modulations in the acoustic ear, enhancing both ongoing and onset ITDs. Also, with CI users in mind, Monaghan and Seeber (2016) proposed an algorithm that identifies local troughs in the envelope and sets these troughs to zero, thus increasing the steepness of local peaks. With this algorithm, they showed improved ITD sensitivity for vocoded speech. It is our hope that by establishing the relationship between temporal fluctuations in speech and binaural sampling, we can motivate the design of novel envelope-based enhancement algorithms designed to improve the delivery of binaural cues. Such algorithms may be particularly useful in multi-source listening environments where binaural cues can yield substantial improvements in speech intelligibility.

It should be noted though that when translating to real-world applications, temporal weighting of both ITDs and interaural level differences (ILDs) should be considered: individually since both cues can provide binaural benefit, and in combination since they co-occur in natural environments. Since CI users have limited access to ITD cues, understanding the temporal weighting of ILD cues in speech is particularly important for these individuals. Future studies will investigate temporal weighting functions for more realistically spatialized speech.

V. SUMMARY

We show that TWFs can be obtained for speech stimuli with a high degree of temporal resolution. In contrast to Baltzell *et al.* (2020), we show that weights are not always highest at the onset of a word, but that instead, weights tend to follow ongoing changes in the acoustic envelope. Model predictions were generated based on the steepness of rising portions of the speech envelope (*Env*), and on simulated auditory-nerve representations (*AuN*), and while both models performed well for certain words, neither could account for the TWFs across all four words.

ACKNOWLEDGMENTS

This work was funded by NIH-NIDCD Award No. DC015760. We would like to thank Gerald Kidd, Chris Mason, and Steve Colburn for their support and many helpful discussions.

[1] Since these distributions were not adjusted such that the ITD cues in each bin (and for each listener) were equally salient (on average), the observed TWFs reflect a combination of ITD sensitivity and effects of temporal position. Because of this, weights should be interpreted as the extent to which equal physical ITDs in each bin contribute to the overall lateralization percept.

[2] There is a conceptual equivalence between "hierarchical" Bayesian models and certain "mixed-effects" frequentist models. Indeed, the "brms" package explicitly follows the syntax of the popular "lmer" package for mixed-effects models. When a mixed-effects model includes both intercept and slope as random effects, it is essentially a hierarchical model, with population-level means estimated as fixed effects.

[3] See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0005934 for intensity-based model predictions.

[4] For unmodulated high-rate click trains, the temporal envelope is essentially flat, and individual clicks should be considered as constituting the TFS. Slight temporal jitter/aperiodicity introduced in these click trains is conceptually equivalent to deviations in the TFS of speech.

Ahrens, A., Joshi, S. N., and Epp, B. (**2020**). "Perceptual weighting of binaural lateralization cues across frequency bands," J. Assoc. Res. Otolaryngol. **21**, 485–496.

Apoux, F., Crouzet, O., and Lorenzi, C. (**2001**). "Temporal envelope expansion of speech in noise for normal-hearing and hearing-impaired listeners: Effects on identification performance and response times," Hear. Res. **153**, 123–131.

Apoux, F., Tribut, N., Debruille, X., and Lorenzi, C. (**2004**). "Identification of envelope-expanded sentences in normal-hearing and hearing-impaired listeners," Hear. Res. **189**, 13–24.

Baltzell, L. S., Cho, A. Y., Swaminathan, J., and Best, V. (**2020**). "Spectro-temporal weighting of interaural time differences in speech," J. Acoust. Soc. Am. **147**(6), 3883–3894.

Berg, B. G. (**1990**). "Observer efficiency and weights in a multiple observation task," J. Acoust. Soc. Am. **88**(1), 149–158.

Best, V., Goupell, M. J., and Colburn, H. S. (**2021**). "Binaural hearing and across-Channel processing," in *Binaural Hearing* (Springer, New York), Vol. 73, pp. 181–208.

Bilsen, F. A., and Raatgever, J. (**1973**). "Spectral dominance in binaural lateralization," Acustica **28**, 131–132.

Brown, A. D., and Stecker, G. C. (**2011**). "Temporal weighting functions for interaural time and level differences. II. The effect of binaurally synchronous temporal jitter," J. Acoust. Soc. Am. **129**(1), 293–300.

Brown, A. D., Stecker, G. C., and Tollin, D. J. (**2015**). "The precedence effect in sound localization," J. Assoc. Res. Otolaryngol. **16**, 1–28.

Bruce, I. C., Erfani, Y., and Zilany, M. S. A. (**2018**). "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites," Hear. Res. **360**, 40–54.

Bürkner, P.-C. (**2018**). "Advanced Bayesian multilevel modeling with the R package BRMS," The R J. **10**(1), 395–411.

Clifton, R. K., and Freyman, R. L. (**1997**). "The precedence effect: Beyond echo suppression," in *Binaural and Spatial Hearing in Real and Virtual*

Environments, edited by R. H. Gilkey and T. R. Anderson (Lawrence Earlbaum, Mahwah, NJ), pp. 233–255.

Culling, J. F., and Colburn, H. S. (**2000**). “Binaural sluggishness in the perception of tone sequences and speech in noise,” J. Acoust. Soc. Am. **107**(1), 517–527.

Desloge, J. G., Reed, C. M., Braida, L. D., Perez, Z. D., and D’Aquila, L. A. (**2017**). “Masking release for hearing-impaired listeners: The effect of increased audibility through reduction of amplitude variability,” J. Acoust. Soc. Am. **141**(6), 4452–4465.

Dietz, M., Ewert, S. D., and Hohmann, V. (**2011**). “Auditory model based direction estimation of concurrent speakers from binaural signals,” Speech Commun. **53**(5), 592–605.

Dietz, M., Marquardt, T., Salminen, N. H., and McAlpine, D. (**2013**). “Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds,” Proc. Natl. Acad. Sci. U.S.A. **110**(37), 15151–15156.

Fitzpatrick, D. C., Kuwada, S., Kim, D. O., Parham, K., and Batra, R. (**1999**). “Responses of neurons to click-pairs as simulated echoes: Auditory nerve to auditory cortex,” J. Acoust. Soc. Am. **106**(6), 3460–3472.

Francart, T., Lenssen, A., and Wouters, J. (**2014**). “Modulation enhancement in the electrical signal improves perception of interaural time differences with bimodal stimulation,” J. Assoc. Res. Otolaryngol. **15**(4), 633–647.

Freyman, R. L., Zurek, P. M., Balakrishnan, U., and Chiang, Y. C. (**1997**). “Onset dominance in lateralization,” J. Acoust. Soc. Am. **101**(3), 1649–1659.

Hafter, E. R., and Buell, T. N. (**1990**). “Restarting the adapted binaural system,” J. Acoust. Soc. Am. **88**(2), 806–812.

Hafter, E. R., and Dye, R. H. (**1983**). “Detection of interaural differences of time in trains of high-frequency clicks as a function of interclick interval and number,” J. Acoust. Soc. Am. **73**(2), 644–651.

Hauth, C. F., and Brand, T. (**2018**). “Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences,” Trends Hear. **22**, 233121651775354.

Hu, H., Ewert, S. D., McAlpine, D., and Dietz, M. (**2017**). “Differences in the temporal course of interaural time difference sensitivity between acoustic and electric hearing in amplitude modulated stimuli,” J. Acoust. Soci. Am. **141**(3), 1862–1873.

Kidd, G., Best, V., and Mason, C. R. (**2008**). “Listening to every other word: Examining the strength of linkage variables in forming streams of speech,” J. Acoust. Soc. Am. **124**(6), 3793–3802.

Klein-Hennig, M., Dietz, M., Hohmann, V., and Ewert, S. D. (**2011**). “The influence of different segments of the ongoing envelope on sensitivity to interaural time delays,” J. Acoust. Soc. Am. **129**(6), 3856–3872.

Koning, R., and Wouters, J. (**2012**). “The potential of onset enhancement for increased speech intelligibility in auditory prostheses,” J. Acoust. Soc. Am. **132**(4), 2569–2581.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). “The precedence effect,” J. Acoust. Soc. Am. **106**(1633), 1633–1654.

Lochner, J. P. A., and Burger, J. F. (**1958**). “The subjective masking of short time delayed echoes, their primary sounds, and their contribution to the intelligibility of speech,” Acoustica **8**(1), 1–10.

Lorenzi, C., Berthommier, F., Apoux, F., and Bacri, N. (**1999**). “Effects of envelope expansion on speech recognition,” Hear. Res. **136**, 131–138.

Louage, D. H. G., van der Heijden, M., and Joris, P. X. (**2004**). “Temporal properties of responses to broadband noise in the auditory nerve,” J. Neurophysiol. **91**(5), 2051–2065.

Lutfi, R. A., Liu, C.-J., and Stoelinga, C. (**2008**). “Level dominance in sound source identification,” J. Acoust. Soc. Am. **124**(6), 3784–3792.

McAlpine, D., Jiang, D., and Palmer, A. R. (**2001**). “A neural code for low-frequency sound localization in mammals,” Nat. Neurosci. **4**(4), 396–401.

Monaghan, J. J. M., and Seeber, B. U. (**2016**). “A method to enhance the use of interaural time differences for cochlear implants in reverberant environments,” J. Acoust. Soc. Am. **140**(2), 1116–1129.

Oberfeld, D. (**2008**). “Does a rhythmic context have an effect on perceptual weights in auditory intensity processing?,” Can. J. Exp. Psychol. **62**(1), 24–32.

Saberi, K. (**1996**). “Observer weighting of interaural delays in filtered impulses,” Percept. Psychophys. **58**(7), 1037–1046.

Schimmel, O., van de Par, S., Breebaart, J., and Kohlrausch, A. (**2008**). “Sound segregation based on temporal envelope structure and binaural cues,” J. Acoust. Soc. Am. **124**(2), 1130–1145.

Slaney, M. (**1993**). “An efficient implementation of the Patterson-Holdsworth auditory filter bank,” Apple Computer Perception Group, Tech. Report No. 35 (Apple, Los Altos, CA).

Stecker, G. C. (**2014**). “Temporal weighting functions for interaural time and level differences. IV. Effects of carrier frequency,” J. Acoust. Soc. Am. **136**(6), 3221–3232.

Stecker, G. C. (**2018**). “Temporal weighting functions for interaural time and level differences. V. Modulated noise carriers,” J. Acoust. Soc. Am. **143**(2), 686–695.

Stecker, G. C., Bernstein, L. R., and Brown, A. D. (**2021**). “Binaural hearing with temporally complex signals,” in *Binaural Hearing*, edited by R. Y. Litovsky, M. J. Goupell, R. R. Fay, and A. N. Popper (Springer, New York), Vol. 73, pp. 145–180.

Stecker, G. C., and Bibee, J. M. (**2014**). “Nonuniform temporal weighting of interaural time differences in 500 Hz tones,” J. Acoust. Soc. Am. **135**(6), 3541–3547.

Stecker, G. C., and Hafter, E. R. (**2002**). “Temporal weighting in sound localization,” J. Acoust. Soc. Am. **112**(3), 1046–1057.

Stern, R. M., and Shear, G. D. (**1996**). “Lateralization and detection of low-frequency binaural stimuli: Effects of distribution of internal delay,” J. Acoust. Soc. Am. **100**(4), 2278–2288.

van der Heijden, K., Rauschecker, J. P., de Gelder, B., and Formisano, E. (**2019**). “Cortical mechanisms of spatial hearing,” Nat. Rev. Neurosci. **20**, 609–623.

Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and Carney, L. H. (**2009**). “A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics,” J. Acoust. Soc. Am. **126**(5), 2390–2412.